大数据时代数字特藏建设探索

——以中国写本文献数字资源库建设为例*

韩松涛[1] 黄 晨[1, 2, 3]
（1. 浙江大学图书馆，杭州 310027；
2. 浙江大学信息资源分析与应用研究中心，杭州 310027；
3. CADAL项目管理中心，杭州 310027）

摘 要：［目的/意义］本文旨在通过中国写本文献数字资源库建设案例，探讨大数据时代数字特藏建设存在的诸多问题。［方法/过程］本文探究了多来源、多形态写本文献资源建设的具体处理方式，明确以学术创新为核心需求的建设目标，以及可持续发展的路径探索等方面，总结数字特藏库建设的模式。［结果/结论］本文认为依托资源核心收藏单位，通过合作进行资源共建，进而完成某类文献的全球整合，并提供方便使用的学术工具，是学术类数字特藏库建设的主要路径与模式。

关键词：数字特藏 写本文献 众包共建
分类号：G256；K877；G250.74
DOI：10.31193/SSAP.J.ISSN.2096-6695.2023.03.04

1 背景综述

传统的特藏数字化工作都是立足馆藏资源、结合学科建设、以项目为抓手构建数据库。这样的构建方式往往存在资源不够完备、技术支撑简陋、后续维护乏力、利用率低下等问题，易形成诸多的资源孤岛。长期以来，图书馆界面对上述困境一直在寻求可行的方案，但似乎一直没有取得有效的突破。

* 本文系全国古籍整理出版规划领导小组办公室2021年度国家古籍数字化工程专项经费资助项目“中国写本文献数字资源库建设（一期）”研究成果之一。
［作者简介］韩松涛（ORCID：0000-0003-4206-4646），男，本科，副研究馆员，古籍特藏部主任，研究方向为古籍碑帖，Email：sthan@zju.edu.cn；黄晨（ORCID：0000-0004-7264-6142），男，硕士，研究馆员，CADAL项目管理中心副主任兼秘书长，研究方向为数字图书馆，Email：huangc@zju.edu.cn（通讯作者）。

2022 年初，在国家古籍数字化工程专项经费资助下，浙江大学启动“中国写本文献数字资源库”建设项目。作为被列入国家古籍数字化十四五规划的重大项目，如何做好顶层设计，最大限度整合全球范围内的中国写本文献资源，贴合学者和公众的使用习惯与需求，形成基于写本文献资源的学术研究空间，是项目组亟需面对的问题。为此，我们专门聘请了浙江大学文科资深教授、敦煌学家张涌泉先生为项目首席学术专家，对资源的搜集、整序、服务利用、研究支持等进行规划指导。项目门户网站（https://xieben.cadal.edu.cn/）于 2022 年 6 月上线,《光明日报》就此进行了专题报道[1]。项目一期于 2022 年 11 月通过了全国古籍整理出版规划领导小组办公室组织的验收。

中国写本文献数字资源库的建设在很多方面都不同于传统的特藏库，虽然当前只是完成了项目一期的建设目标，但是建设过程中的理念、规划、组织乃至技术架构都有可供借鉴讨论之处。本文拟以中国写本文献数字资源库项目为例，就大数据时代数字化特藏资源构建进行一些探讨，期望能够提出一种数字人文背景下基于资源的学术研究平台的可持续发展模式，为学界业界同行提供借鉴参考。

2　多来源、多形态资源的众包共建

写本文献是指用软笔、硬笔书写在纸张上的古籍或文字资料，主要流行于东汉至五代时期，是这一时期中华文明传承的主要载体。北宋以后，随着雕版印刷技术的日趋成熟，刻本逐渐取代了写本的主体地位。刻本的流行，既在事实上加速了早期古写本的湮没，也在内容上形成了以契约文书、法律档案、民间宗教写本、帐目便笺及通俗文学作品等为核心的写本系统。清末以来，国内外的科学家和探险者曾先后在甘肃、新疆、陕西一带发现了一些早期的写本文献，包括西汉时期的古地图、晋代的《战国策》《三国志》写本等，但数量有限。

1900 年 6 月 22 日，敦煌莫高窟藏经洞被打开，人们从中发现了大批唐代前后的写本文献，震惊了整个世界。民国时期以后，又有吐鲁番文书、黑水城文献、宋元以来契约文书、明清档案等众多写本文献陆续公诸于世，辉耀世界，写本文献的数量一下充盈起来，其又重新回到世人的视域之中[2]。

2.1　多来源、多形态特藏资源的搜集

写本文献的类型众多，分布广泛，为资源收集带来较大的不确定因素。以敦煌文献、吐鲁番文书和地方文书为例，每一类写本文献分布都具有分散性特征，其图像也以多形态的方式存在。

2.1.1　敦煌文献

敦煌文献主要由中国国家图书馆、英国国家图书馆、法国国家图书馆和俄罗斯科学院东方文献研究所四家机构收藏。其中，中国藏敦煌遗书约 16 000 件，俄藏敦煌文献约 19 000 件（含残片），英藏敦煌文献约 14 000 件，法藏敦煌文献约 8 100 件。法藏、英藏部分的内容最为丰富，亦最具研究价值。此外，国内还有多家公私收藏机构收藏，其他国家也有少量收藏，包括日本、美国、德国、丹麦、瑞典等[3]。其图像，早期有黑白胶片，后期有灰度及彩色出版的图书，近期有彩色电子图片，呈现与时代技术相对应的状态。

2.1.2　吐鲁番文书

吐鲁番文书现存约 4 万余件，分散于世界各地。其中中国收藏约 1.2 万余件，分别收藏于新

疆维吾尔自治区博物馆、中国国家博物馆、北京大学图书馆、上海图书馆及辽宁省档案馆等地。国外分别收藏于德国国家图书馆、日本京都龙谷大学图书馆、俄罗斯科学院东方学研究所圣彼得堡分所、芬兰赫尔辛基大学图书馆、英国图书馆、伊斯塔坦布尔大学图书馆、美国普林斯顿大学图书馆等[4]。除一定规模收藏外，零散的收藏更不在少数，如浙江大学图书馆就藏有少量高昌国时期的吐鲁番文书。吐鲁番文书的研究起步较晚，公开发布也较零散，且以出版的图书为主。

2.1.3 地方文书

地方文书的分布更加零散。如较为著名的徽州文书，作为迄今最大宗的民间文献，它是 20 世纪 50 年代在屯溪古籍书店收集、出售古书的过程中被发现的。此后中国科学院等机构开始大量收购徽州文书，开启了民间文书大规模收集的先河[5]。据不完全统计，世界各地收藏机构和个人收藏的徽州文书达 80 余万件，其中中大图书馆收藏约 33.8 万件，数量之巨大、时代之久远、内容之丰富、记录之系统，堪称地方文献收藏典范[6]。但从公开出版的图书看，徽州文书发布只有数万件，大量的徽州文书还主要存藏于各大公藏之中。

再如，浙江地方文书收藏也比较零散。其中浙江师范大学收藏约 10 万件，分别收藏于该校图书馆和出土文献中心。浙江大学图书馆收藏近万件，且数量还在不断增长中。浙大城市学院也开始收藏地方文书，数量已有万件以上。这些公藏的地方文书基本都是从民间收集而来。另外公藏机构也藏有一些作为档案流存下来的地方文书。比如，浙江龙泉有一批司法档案，存藏于该市档案馆[7]。又有著名的兰溪鱼鳞图册，收藏于兰溪市档案馆[8]。地方文书也有个人收藏，如个人建的石仓契约博物馆，收藏大批的契约文书。以上可见，地方文书收藏机构众多，整合起来难度较大。同时，除石仓契约出版八千余件，龙泉司法档案、兰溪鱼鳞图册等已经出版之外，其他浙江地方文书尚处于整理的初期，数据库的相关文献搜集工作只能通过多方合作的方式完成。

2.2 特藏资源的整序

如此浩繁分散、形式多样的写本文献，在搜集过程中亟需解决资源组织的规范问题。分类和标引是资源整序的重要方式，中国写本文献数字资源库在建设之初就制定了分类和入库数据的元数据标准。

2.2.1 中国写本文献数字资源库的分类

分类的目的在于观瞻和浏览方便，同时实现事实上的文献相互关联。中国写本文献数字资源库采用地域分类与内容分类两种方式并行施用。

按地域分类，意指某一地新发现、刊布的写本文献明显推动了某一种特色明显的学术研究的出现，更侧重于宏观性、整体性。在原始资源上，写本按地域分类，并考虑资源的全球影响力，暂拟分作吐鲁番文献、敦煌文献、黑水城文献三类。吐鲁番文献，指 19 世纪末以来在新疆吐鲁番地区晋唐古墓葬群中所发现的写本文献，是魏晋六朝纸本文献的主要实物遗存。敦煌文献，主要指敦煌莫高窟藏经洞发现的唐代前后的写本文献以及少量的刻本文献。敦煌文献的抄写时代上起魏晋六朝，下迄宋初，前后跨越六百多年，而以唐五代为主体，前承吐鲁番文书，后接宋元以后刻本及写本文献，是唐五代纸本文献的主要实物遗存。黑水城文献，指在内蒙古额济纳旗黑水城遗址发现的纸质写本以及少量的刻本文献。黑水城文献前承敦煌文献，其抄写、刻印年代为北宋、辽、金、西夏、元、北元时期，以西夏文和汉文文献为主，内容涉及传统经史子集类图书，以及佛经、道经、契约文书、官方档案等，是研究该时期特别是西夏王朝的珍贵资料。

按内容分类，中国写本文献数字资源库暂拟分作地方文书、明清档案、民间宗教写本、民间戏曲小说写本等四类。地方文书，与官府档案文书相对而言，指近一个世纪以来陆续发现的宋至民国时期以手写为主的民间文书。明清档案，是指明清宫廷和各级政府部门的档案，现存的明清档案约有二千万件之巨。民间宗教写本，指流行于古代百姓阶层的数十种民间教派的典籍抄写本，具有民俗学、宗教学等多重价值，是中华传统文化的一个重要组成部分。民间戏曲小说写本，主要是指通过手写方式得以保存流传的戏曲脚本、戏曲乐谱、戏曲理论、说唱文本、杂剧故事等文献，与刻印本形成了对应的类别。

以上写本文献，从时间轴线看，约从魏晋六朝直至清代，这些写本文献成为连续不断地反映上述历史时期的文学、历史、思想、民俗、宗教等的重要资源。上述分类是一级分类，在一级分类下还有二级、三级分类，比如地方文书包含了 13 个二级类目和 47 个三级类目。

2.2.2　中国写本文献数字资源库的标引

资源标引主要采用元数据的方式，初步原则是希望与都柏林核心元数据标准能够相互映射。以写本元数据为例，将都柏林核心元数据 15 项元素分为内容、形态、收藏三大类，分别对应相关的写本内容。同时对部分项目作了扩展，比如将题名扩展为自拟题名和原题名两类。这是因为大多数写本不存在原题名，故对写本的描述只能自拟，同时设定了自拟题名的规则。又如，时间扩展为三项，分别是朝代、中国纪年（年号）、公元纪年。这样有三个作用，一是能够进行客观记述，二是能够作为精确的检索点，三是可以用作导航使用。

资源标引还可以采用开放标签的方式。不过，单维度的标签在学术资源的分类中无法发挥较好的作用。不同维度的标签在同一个“平面”内使用，无法替代体系分类法的内容分类效果。我们关注到阮冈纳赞的《冒号分类法》第七版，其实质是一部“分面组配式分类法”。“标签和‘分面组配式分类法’本质上讲，有着很大的相似度，‘分面组配式分类法’其实是一个多维度（或称多个分面）的标签，利用维度的概念去规范标签，而又让标签在维度内保证它的自由度，是一种让标签发挥其学术分类作用的最主要方法。”[9] 多维度标签也将是中国写本文献数字资源库开放的资源标引方式之一。

2.3　众包共建的可持续发展

任何一个收藏单位很难穷尽某一领域的文献资源收藏。写本文献来源广泛，形态各异，标引繁杂，这样一个资源库的建设，如果单纯依靠单馆建设，无疑只能是一种撮要示例式的特藏库，没有太大的实用价值。立足馆藏优势，共建共享是重要的途径。其他馆藏尤其是分散在全球的资源如何参与共建，从实物到图片、出版物，乃至影像资源，基于赛博空间的资源整合与无缝连接，是特藏众包建设的关键问题。

为此我们在特藏资源平台构建的过程中加入了开放众包的功能要求。一方面系统通过后台爬虫和前台编辑，允许来自于不同公私收藏的资源整合，注明来源和明确权限控制。我们设计了可扩展的元数据框架，随时可以在线增加同种资源不同来源的描述，也允许对文本、图片、URL 乃至研究论文的添加。另一方面系统还允许用户上传图片、添加注释、纠正元数据等，上述这些在经过管理员审核后成为资源库的新内容。另外，我们还在尝试开发 API 和接口标准，以便有合作意向的机构可以通过接口实现关联数据，最大可能地向用户揭示更多的资源。

实际上，研究是一个不断利用前人成果的累积过程，可以看作是众包的原始形态，而互联网使这种历时性的过程能够以同时与历时两种方式展开，特藏建设引入众包机制可以让不同机构、个人进行在线协作，不断提升资源完整度，完善资源描述与组织框架，紧密联结资源服务与研究，有效解决特藏资源库建设的可持续发展问题。

3　从资源服务到研究支撑

传统的文献服务专注于可发现与可获得，无论是在地还是在线的方式都依然是外化于用户的学习和研究过程的。如何通过平台的技术功能模块支持和增值资源服务，使其真正嵌入用户的学术生命周期，是新一代特藏资源库的重要标志。

3.1　虚拟整合与关联缀合

由于资源的多来源特征，那么多来源展示与比对在单个文献著录基础上的多维度关联性揭示等功能，便成为中国写本文献数字资源库的重要需求。资源库不仅满足于文献简单呈现，而且提供一些工具让用户执行具有可操作性，既让资源关联性得到揭示，也让用户的发现和学术成果能够得到展现，从而增加用户的粘性。

3.1.1　多来源数据展示与多资源关联揭示

文献以件为单位著录，而单件文献往往有多个复本来源。以敦煌文献为例，同一件敦煌卷子，因为时代变革、技术发展等原因，产生了多种不同的再生性保护成果，如早期的缩微胶卷、图书出版等大多为黑白的图像，以及后期技术的发展后产生的彩色的图像，主要包括图书和数字化扫描图片。一般认为，后期的技术更为先进，其图像清晰度更好，只要保留最后一次的扫描成果就可以了。然而不同再生性保护成果之间前后会存在较长的时间差距，像敦煌文献以唐代的写本为主，出土最晚年代为宋代，至今已超过了千年，每进行一次拍摄或扫描，都会对文献造成损伤。虽然早期的复制技术比较落后，但早期写本文献的完整性却是最好的。同时每次对文献的复制都有可能发生一些“事故”，比如扫描对边角位置的遗漏、对反面只有个别符号的情况不加关注等。有鉴于此，对于收集完整相关写本文献的要求来说，收集完所有编号文献并不等于收集全了，而是要对于历次的复制成果都予以收录，才是数据完整性的保障。中国写本文献数字资源库对此进行了搜集和标引，并支持多来源展示，提供不同来源图像的对比功能（见图 1）。考虑到版权和可获得性等因素，有些数据仅通过 URL 访问方式展示。

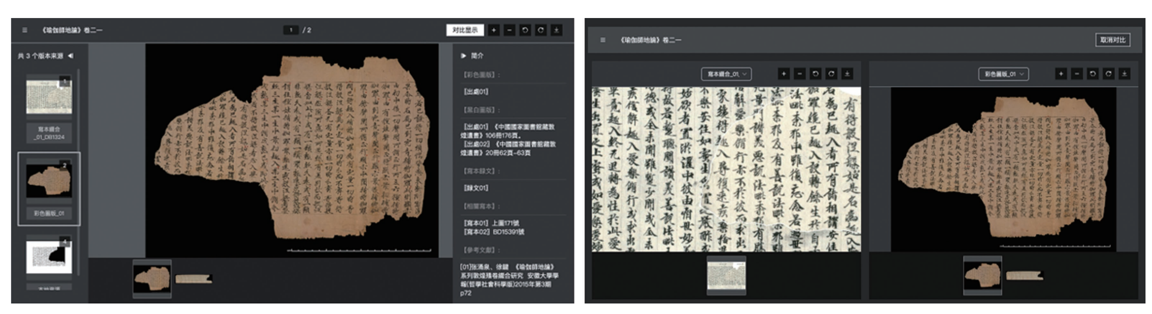


图 1　多来源版本展示与比对

此外，多个写本文献之间存在着关联，若以单件形式在数据库中呈现，是无法充分揭示其所有内涵的，必需同时揭示其关联性，才能使用户更好地使用资源。我们以单件著录，并提供组合功能，通过虚拟整合的方式让单件著录的文献加入群组来满足两者兼顾的要求。为此，我们设定了两个虚拟专题：关联与群组。

关联专题，是指将同一种文献或同一类文献的不同时期写本按一定的形式放在一起，建立相互之间的关联，有助于用户对比研究。例如《论语》算是唐代的“教材”之一，在敦煌与吐鲁番出土的文献中都有当时学郎抄写的不同写本，至少包括了《论语郑氏注》《论语集解》《论语义疏》《论语音义》等多个《论语》的注本。在对原始写本扫描收录的基础上，完全可以建立一个“《论语》关联专题”，将敦煌、吐鲁番发现的所有写本一并呈现，并建立相互之间的联系，必然能够最大化发挥这些写本的学术价值，弥补现存史籍记载的缺漏，更能够向用户呈现出专题化、多样化的关联效果。

群组专题，主要是针对契约文书以及与契约文书类似的一些写本的特殊情况。契约文书虽以每一件作为著录单位，但往往以一个家族一批的形式出现，称为“归户”，只有归户的契约才具有较高的研究价值。从著录标准看，单单是“题名”一项，多个文书就会出现多个相似的题名，数百张在一起则必然会出现无法著录的情况。如果另取一个总题名著录，则与客观著录的原则相违背。因此，借鉴学术界“归户整理法”的思维，对于这一类的写本可以通过群组专题的形式予以呈现。事实上，这些留存下来的文书大都是原始的凭据、字据等记录，它们曾经与文书主人的生产、生活、社会交往、情感世界等紧密相关，同属一个主体，彼此之间也相互关联，由此构成了一个连续性的整体，体现出一种内在的归属性。借助“归户整理法”思维，建立“归户”的群组专题，把同一批发现的同一个家族文书做成一个群组，将与之相关联的所有写本联系起来，既能够单独呈现每一件写本的客观情况，便于检索，又能够通过组建群组专题，揭示文书之间的组合关系与整体面貌。

总的来说，以件为单位著录，采用随机可组专题的虚拟整合形式揭示写本文献之间的关系，是中国写本文献数字资源库建设的较佳方案。同时，前文提到的多维度标签，即通过维度标签的形式动态聚类，也是解决关联性的方式之一。

3.1.2 一种特殊的虚拟整合——缀合

缀合专题，主要应用于敦煌文献。敦煌卷子历史上由于各种原因，造成了整卷文献撕成多段或较小碎片的情况。原本同属一件、如今被人为撕裂为多件的敦煌文书，中国写本文献数字资源库是以所藏地编号来著录的，这样会将这些原属于同一文书的残片分散著录。所以，平台支持数据库建设者及用户以组建专题的形式，将相关的多件写本组成一个缀合专题，并作出提示说明。项目首席学术专家张涌泉教授一直从事敦煌文献的缀合，缀合专题的建设现在主要是利用其研究成果，将已知佚散的残件文献关联到一个主题（见图 2）。

随着人工智能技术的发展，智慧古籍平台的探索建设也开始进入实施阶段，其过程主要是通过学者进行总体框架和相关内容的制定，再利用一定的技术手段呈现。我们正在与计算机领域的专家合作，尝试使用机器学习的方式，通过形状分析和笔迹比对进行文献碎片的缀合，利用人机交互快速实现资源的重组和完善，未来有希望利用 AI 技术快速实现系统内资源的缀合，方便用户的学习与研究。



图 2 英国收藏的几件敦煌写本缀合效果示意图

3.2 基于资源空间的学术社区

传统特藏库提供的是一种单向度的文献服务，用户在其中搜索、获取需要的资源，然后回到工作平台继续自己的研究。如果能够将资源空间和研究空间打通甚至融合，用户在一个空间中完成资源的搜索、标引、注释、重组，形成个人研究方向的知识空间，必然可以极大提高学习与研究的效率。同时，这样的知识空间也可以不断增益特藏库的资源空间，使之成为一个自生长的特藏资源库，这是新一代特藏资源平台建设的重要发展方向。

我们的思路是在特藏资源平台引入一系列的工具，比如基于历史地理的 GIS 系统，允许用户通过个人空间自由标引标注，从时空维度组织与整合平台资源及自有资源；提供通用知识库，包括年号纪年转换，简繁体转换，职官、地名、人名、名物等查询，多语种翻译等等；提供可视化工具包，方便内容的图表化展示……这些工具将资源无缝地融入学习与研究环境，支持资源的多维度展示以促进学习与研究发现。

为此，我们也在积极考察国内外同行的解决方案。在全球范围内，使用 URI，采用 W3C 开放标准，关联其他数据以形成虚拟整合与研究空间的项目有多个，比如 DDB、BNB、LC、DPLA、Getty 博物馆、ResearchSpace、上海图书馆等[10]。以 ResearchSpace 为例，该项目由不列颠博物馆与牛津大学共同开发，致力于通过关联数据汇聚全球在线资源形成特藏知识空间，是一个透明穿梭于特定主题的全球资源平台。网站对自己的介绍是“连接、关联、语境知识表达”（Connect, communicate and represent knowledge with context）[11]。大学数字图书馆国际合作计划（CADAL）项目从 2017 年起与牛津大学开展合作，成功地将中国传统绘画、音乐文物两类资源与 ResearchSpace 系统进行了链接。我们希望借鉴 ResearchSpace 的框架，逐步将特藏资源平台构建成基于资源空间的学术社区。

4 可持续发展与利用

资源库如何成为一个“生长着的有机体”，除了建设者不断更新内容外，用户参与也是主要方式之一。当平台从文献服务到知识服务、再到出版服务，甚至成为用户学习研究空间的一个赛博研究室的时候，用户创建的大量内容，一方面作为学术档案为学习研究提供支撑，另一方面除了作为学术论文和专著在平台展示之外，还可以进一步拓展平台特藏资源的学术空间。

我们认为有几个方式可以考虑：一是网站拥有自己独特的资源，让用户不得不来；二是具有整合性，使某一类资源具有完整性或相对来说的最大集合；三是共建共享，推进多方合作；四是网站提供一系列学术工具，形成基于资源的研究支撑平台。

中国写本文献数字资源库就上述四个方面进行了规划，并已经初步进行了建设。

4.1 依托浙江大学的独特写本资源建设

浙江大学收藏的写本文献主要包括敦煌吐鲁番文书、地方文书、宗教写本等。其中吐鲁番文书是一个非常独特的收藏，共计 20 件，其中大部分保留冥器的原始形态，没有被拆解，这在全球范围都是比较珍贵的遗存。项目组拟就浙江大学收藏的吐鲁番文书进行三维数字化扫描后再进行拆解。

浙江大学图书馆还收藏有一批以浙江为主、兼有部分相邻省份的地方文书，主要地域范围涉及浙江、福建两地，包括丽水云和县、松阳县、景宁县及温州泰顺乡等，现收藏共计约 1.5 万张 / 册 / 件。其中以“浙江丽水庆元举水荐坑吴姓民间文书”这一组最具价值，该批地方文书共计 1200 余件，为吴姓家族从明代中期到民国时期的经济往来、民间活动等文书，具有很高的研究和利用价值。20 世纪以来，随着史学研究视角下移，利用民间历史文献成为潮流。地方文书对于研究乡村社会组织、地权结构、赋役制度、人民生计、宗教信仰、语言文字等极具史料价值，故而备受学界青睐。从文化保存的意义上看，地方文书是一种记载历史文化的重要载体，与出土文物一样具有文物价值，每一份地方文书都具有独特性与不可复制性。其与古籍刊本不同，多为手写本，往往仅限于内部流通，大多为传世的“孤本”。因此，地方文书文献的搜集与整理，既构成了学术基础，又构成了学术前沿，具有文化保存与学术研究的双重价值。

浙江大学图书馆收藏的宗教写本包括数幅长度 8 至 10 米的佛经长卷，以及部分佛教道教经典及科仪本等抄本，均颇具特色。

通过购藏、捐赠及对已有馆藏的组织整理，上述各类特色文献数量还在不断增长中。

4.2 建设吐鲁番文书发布站点

从数字化历程看，敦煌文献已经有较为成熟的国际敦煌项目（IDP）。IDP 项目于 1994 年正式启动，其核心工作是关于敦煌与新疆出土古文献及文物的修复、编目与保护。随着网络技术的发展，IDP 项目希望通过与收藏机构开展合作，以高质量的数字图像将这些艺术品重新拼合在一起，方便学者和公众在网上获得越来越多的相关信息。

与之相比，当前吐鲁番文书在全球尚没有一个网站进行系统收录，中国写本文献数字资源库将把吐鲁番文书作为一个专题进行重点建设，使之成为整合全球吐鲁番文书的发布中心，推动吐鲁番文书的保护、研究、共享建设，并建立相关的区域中心。

4.3 依托 CADAL 进行共建共享

大学数字图书馆国际合作计划（CADAL）作为教育部建设的公益数字图书馆项目，已经与国内高校及国外数十所知名高校图书馆、国家图书馆建立了广泛的合作。中国写本文献数字资源库将依托 CADAL 项目开展国内高校和国外写本收藏机构的合作推进项目建设。以存世数百万件的地方文书为例，项目门户网站拟收录已公开发表的文书，并联合大宗收藏机构共同发布。

4.4 加强写本文献学术专著和工具书建设

以敦煌文献为首要建设方向。一是将每件敦煌卷子的缀合信息、学术研究成果分列在文献之后，提供学术参考。二是对敦煌学的工具书进行重点建设，使项目门户网站的学术研究和参考作用得到强化。

5 结论与建议

以“中国写本文献数字资源库”建设项目为案例，我们就特藏的构建、组织、服务和利用进行了系统梳理，做了一些有益的尝试，更多的是进行了规划和思考。在此过程中，我们发现特藏资源建设在来源、形态、组织和呈现方面都需要观念与技术实现突破，局限于一馆一地，仅针对一两种介质的资源，完全依赖于规范的分类原则和受控词表的标引，通过关键词搜索提供访问的网站，都将使特藏资源库成为固步自封的孤岛，逐渐淹没在互联网的泡沫中。

大数据时代，资源的揭示更为充分，网络条件能够支撑多样化的分工合作，数字资源的独占性不像纸本资源那样突出，完全可以通过众包方式，元数据与内容相结合，可发现与可获得相平衡，构建比传统数据库远为完备、使用体验更佳、利用率与可持续良性循环的数字特藏库。

有鉴于此，我们将中国写本文献数字资源库的建设定位于面向学者和公众为主的高质量用户的基于资源研究与学习的平台。一是立足本地放眼全球，数字化整合尽量完整的主题资源；二是提供众包标引工具，鼓励用户进行多维度标引从而产生多向度的关联；三是开放学术空间，满足用户阅读、学习、资源聚合乃至知识创造的需求。这些设想有些已经在项目的一期建设中实现，并且得到了很好的使用反馈，有些则还有待于在项目的后续建设中持续完善和深化。我们希望通过打造这样一个全新平台，吸引不同领域的学习和研究的用户，既能够通过他们的使用体验来不断提升平台可用性，又能够让他们参与到特藏库的内容生成，从而发展出一个自生长自组织的特藏构建模式，使其成为数字人文研究与新文科实验室建设的典范。

【参考文献】

[1] 杜羽，陆健. “中国写本文献数字资源库”发布［N］. 光明日报，2022-6-26（4）.

[2] 张涌泉. 写本文献：中华文明传承的重要载体，从考古看中国［M］. 北京：中华书局，2022.

[3] 赵彦昌，李晓光. 论敦煌文献流失海外的原因、经过及具体分布［J］. 辽宁省博物馆馆刊，2012（00）：386-404.

[4] 王小苹. 吐鲁番文书流失海外的实证研究［J］. 大众文艺，2016（2）：273-274.

[5] 杨培娜，申斌. 走向民间历史文献学——20世纪民间文献搜集整理方法的演进历程［J］. 中山大学学报（社会科学版），2014，54（5）：71-80.

［6］王蕾. 徽州文书、徽学研究与数字人文［J］. 图书馆论坛，2016，36（9）：1-4.

［7］尹伟琴. 论民国时期基层法院判决依据的多样性——以浙江龙泉祭田纠纷司法档案为例［J］. 浙江社会科学，2010（5）：92-96.

［8］高超群. 新书《兰溪鱼鳞图册合集》简介［J］. 中国经济史研究，2023（1）：103.

［9］韩松涛. 数字图书馆分类法新论［J］. 图书馆杂志，2011，30（10）：36-39.

［10］张喆昱，张磊. 记忆机构的开放数据建设和数字化服务转型［J］. 图书馆论坛，2020，40（5）：21-26.

［11］ResearchSpace［EB/OL］.［2023-6-25］. https://researchspace.org/.

Exploration of Digital Collection Construction in the Era of Big Data: A Case Study of the Construction of Chinese Manuscript Digital Collection

Han Songtao[1] Huang Chen[1,2,3]

(1. Zhejiang University Library, Hangzhou 310027, China;
2. Center for Information Resources Analysis & Application of Zhejiang University, Hangzhou 310027, China;
3. Administration Center for China Academic Digital Associative Library, Hangzhou 310027, China)

Abstract: ［**Purpose/significance**］Focusing on a case study of the construction of the digital database of Chinese manuscript documents, this paper aims to discuss the problems existing in the construction of digital collection in the age of big data. ［**Method/process**］It probes into the concrete methods in the construction of multi-source and multi-form manuscript document resources and makes it clear that the goal of construction is to meet the core requirement of academic research. It also explores aspects including the path of sustainable development, making a conclusion on the mode of digital collection database construction. ［**Result/conclusion**］To conclude, the paper maintains that the main path and mode of the construction of an academic digital collection database is to achieve core-construction through the collaboration of core resource collection entities and to provide academic tools easy to use, therefore completing the integration of certain documents at a global level.

Keywords: Digital special collection; Manuscript documents; Crowdsourced and co-construction

（本文责编：魏　进）